Rare germline copy number variants (CNVs) and breast cancer risk

Authors

Joe Dennis[1], Jonathan P. Tyrer[2], Logan C. Walker[3], Kyriaki Michailidou[1, 4, 5], Leila Dorling[1], Manjeet K. Bolla[1], Qin Wang[1], Thomas U. Ahearn[6], Irene L. Andrulis[7, 8], Hoda Anton-Culver[9], Natalia N. Antonenkova[10], Volker Arndt[11], Kristan J. Aronson[12], Laura E. Beane Freeman[6], Matthias W. Beckmann[13], Sabine Behrens[14], Javier Benitez[15, 16], Marina Bermisheva[17], Natalia V. Bogdanova[10, 18, 19], Stig E. Bojesen[20-22], Hermann Brenner[11, 23, 24], Jose E. Castelao[25], Jenny Chang-Claude[14, 26], Georgia Chenevix-Trench[27], Christine L. Clarke[28], NBCS Collaborators*, J. Margriet Collée[29], CTS Consortium†, Fergus J. Couch[30], Angela Cox[31], Simon S. Cross[32], Kamila Czene[33], Peter Devilee[34, 35], Thilo Dörk[19], Laure Dossus[36], A. Heather Eliassen [37, 38], Mikael Eriksson[33], D. Gareth Evans[39, 40], Peter A. Fasching[13, 41], Jonine Figueroa[6, 42, 43], Olivia Fletcher[44], Henrik Flyger[45], Lin Fritschi[46], Marike Gabrielson[33], Manuela Gago-Dominguez[47, 48], Montserrat García-Closas[6], Graham G. Giles[49-51], Anna González-Neira[16], Pascal Guénel[52], Eric Hahnen[53, 54], Christopher A. Haiman[55], Per Hall[33, 56], Antoinette Hollestelle[57], Reiner Hoppe[58, 59], John L. Hopper[50], Anthony Howell[60], ABCTB Investigators‡, kConFab Investigators§, Agnes Jager[57], Anna Jakubowska[61, 62], Esther M. John[63, 64], Nichola Johnson[44], Michael E. Jones[65], Audrey Jung[14], Rudolf Kaaks[14], Renske Keeman[66], Elza Khusnutdinova[17, 67], Cari M. Kitahara[68], Yon-Dschun Ko[69], Veli-Matti Kosma[70-72], Stella Koutros[6], Peter Kraft[38, 73], Vessela N. Kristensen[74, 75], Katerina Kubelka-Sabit[76], Allison W. Kurian[63, 64], James V. Lacey[77, 78], Diether Lambrechts[79, 80], Nicole L. Larson[81], Martha Linet[68], Alicja Lukomska[61], Arto Mannermaa[70-72], Siranoush Manoukian[82], Sara Margolin[56, 83], Dimitrios Mavroudis[84], Roger L. Milne[49-51], Taru A. Muranen[85], Rachel A. Murphy[86, 87], Heli Nevanlinna[85], Janet E. Olson[81], Håkan Olsson[88], Tjoung-Won Park-Simon[19], Charles M. Perou[89], Paolo Peterlongo[90], Dijana Plaseska-Karanfilska[91], Katri Pylkäs[92, 93], Gad Rennert[94], Emmanouil Saloustros[95], Dale P. Sandler[96], Elinor J. Sawyer[97], Marjanka K. Schmidt[66, 98], Rita K. Schmutzler[53, 54, 99], Rana Shibli[94], Ann Smeets[100], Penny Soucy[101], Melissa C. Southey[49, 51, 102], Anthony J. Swerdlow[65, 103], Rulla M. Tamimi[38, 104], Jack A. Taylor[96, 105], Lauren R. Teras[106], Mary Beth Terry[107], Ian Tomlinson[108, 109], Melissa A. Troester[110], Thérèse Truong[52], Celine M. Vachon[111], Camilla Wendt[83], Robert Winqvist[92, 93], Alicja Wolk[112, 113], Xiaohong R. Yang[6], Wei Zheng[114], Argyrios Ziogas[9], Jacques Simard[101], Alison M. Dunning[2], Paul D.P. Pharoah[1, 2], Douglas F. Easton[1, 2].

[1] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
[2] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.
[3] Department of Pathology and Biomedical Science, University of Otago, Christchurch, New Zealand.
[4] Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus.
[5] Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus.
[6] Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA.
[7] Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada.
[8] Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada.
[9] Department of Medicine, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA, USA.

[10] N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus.
[11] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[12] Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON, Canada.
[13] Department of Gynecology and Obstetrics, Comprehensive Cancer Center Erlangen-EMN, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg (FAU), Erlangen, Germany.
[14] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
[15] Biomedical Network on Rare Diseases (CIBERER), Madrid, Spain.
[16] Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain.
[17] Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia.
[18] Department of Radiation Oncology, Hannover Medical School, Hannover, Germany.
[19] Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
[20] Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[21] Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[22] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
[23] Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.
[24] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
[25] Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain.
[26] Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
[27] Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia.
[28] Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia.
[29] Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands.
[30] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.
[31] Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK.
[32] Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK.
[33] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
[34] Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
[35] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.
[36] Nutrition and Metabolism Section, International Agency for Research on Cancer (IARC-WHO), Lyon, France.
[37] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.
[38] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[39] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK.
[40] North West Genomics Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.
[41] David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, USA.
[42] Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh, Edinburgh, UK.
[43] Cancer Research UK Edinburgh Centre, The University of Edinburgh, Edinburgh, UK.
[44] The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK.
[45] Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
[46] School of Public Health, Curtin University, Perth, Western Australia, Australia.
[47] Fundación Pública Galega de Medicina Xenómica, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain.
[48] Moores Cancer Center, University of California San Diego, La Jolla, CA, USA.
[49] Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia.
[50] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia.
[51] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia.
[52] Center for Research in Epidemiology and Population Health (CESP), Team Exposome and Heredity, INSERM, University Paris-Saclay, Villejuif, France.
[53] Center for Familial Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[54] Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[55] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.
[56] Department of Oncology, Södersjukhuset, Stockholm, Sweden.
[57] Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands.
[58] Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.
[59] University of Tübingen, Tübingen, Germany.
[60] Division of Cancer Sciences, University of Manchester, Manchester, UK.
[61] Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland.
[62] Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland.
[63] Department of Epidemiology & Population Health, Stanford University School of Medicine, Stanford, CA, USA.
[64] Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA.
[65] Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
[66] Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands.
[67] Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia.
[68] Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA.

[69] Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany.
[70] Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.
[71] Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland.
[72] Biobank of Eastern Finland, Kuopio University Hospital, Kuopio, Finland.
[73] Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
[74] Department of Medical Genetics, Oslo University Hospital and University of Oslo, Oslo, Norway.
[75] Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.
[76] Department of Histopathology and Cytology, Clinical Hospital Acibadem Sistina, Skopje, Republic of North Macedonia.
[77] Department of Computational and Quantitative Medicine, City of Hope, Duarte, CA, USA.
[78] City of Hope Comprehensive Cancer Center, City of Hope, Duarte, CA, USA.
[79] VIB Center for Cancer Biology, Leuven, Belgium.
[80] Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium.
[81] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.
[82] Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy.
[83] Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden.
[84] Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece.
[85] Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland.
[86] School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada.
[87] Cancer Control Research, BC Cancer, Vancouver, BC, Canada.
[88] Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden.
[89] Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
[90] Genome Diagnostics Program, IFOM - the FIRC Institute of Molecular Oncology, Milan, Italy.
[91] Research Centre for Genetic Engineering and Biotechnology 'Georgi D. Efremov', MASA, Skopje, Republic of North Macedonia.
[92] Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland.
[93] Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland.
[94] Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel.
[95] Department of Oncology, University Hospital of Larissa, Larissa, Greece.
[96] Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.
[97] School of Cancer & Pharmaceutical Sciences, Comprehensive Cancer Centre, Guy's Campus, King's College London, London, UK.
[98] Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands.
[99] Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany.
[100] Department of Surgical Oncology, University Hospitals Leuven, Leuven, Belgium.
[101] Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval Research Center, Québec City, QC, Canada.

[102] Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia.
[103] Division of Breast Cancer Research, The Institute of Cancer Research, London, UK.
[104] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA.
[105] Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.
[106] Department of Population Science, American Cancer Society, Atlanta, GA, USA.
[107] Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA.
[108] Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK.
[109] Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK.
[110] Department of Epidemiology, Gillings School of Global Public Health and UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.
[111] Department of Quantitative Health Sciences, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA.
[112] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
[113] Department of Surgical Sciences, Uppsala University, Uppsala, Sweden.
[114] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA.

* NBCS collaborators are listed at the end of the paper.

† Members of the CTS Consortium are listed at the end of the paper.

‡ ACTB investigators are listed at the end of the paper.

§ kConFab investigators are listed at the end of the paper.

Corresponding Author:

Joe Dennis, Centre for Cancer Genetic Epidemiology, University of Cambridge

jgd29@cam.ac.uk

Abstract

Germline copy number variants (CNVs) are pervasive in the human genome but potential disease associations with rare CNVs have not been comprehensively assessed in large datasets. We analysed rare CNVs in genes and non-coding regions for 86,788 breast cancer cases and 76,122 controls of European ancestry with genome-wide array data.

Gene burden tests detected the strongest association for deletions in *BRCA1* (P= 3.7E-18). Nine other genes were associated with a p-value < 0.01 including known susceptibility genes *CHEK2* (P= 0.0008), *ATM* (P= 0.002) and *BRCA2* (P= 0.008). Outside the known genes we

detected associations with p-values < 0.001 for either overall or subtype-specific breast cancer at nine deletion regions and four duplication regions. Three of the deletion regions were in established common susceptibility loci.

To the best of our knowledge, this is the first genome-wide analysis of rare CNVs in a large breast cancer case-control dataset. We detected associations with exonic deletions in established breast cancer susceptibility genes. We also detected suggestive associations with non-coding CNVs in known and novel loci with large effects sizes. Larger sample sizes will be required to reach robust levels of statistical significance.

Introduction

Copy number variants (CNVs) are pervasive in the human genome but are more challenging to detect with current technologies than single nucleotide variants (SNVs). Recent comprehensive sequencing projects [1,2] have characterised CNVs in large sample sets. The gnomAD project identified a median of 3,505 deletions and 723 duplications covering more than 50 base pairs per genome. Most deletions and duplications tend to be rare with longer variants tending to be rarer, suggesting negative selection against these variants. At the population level the 1000 Genomes project has mapped a large proportion of inherited CNVs [3] and observed that 65% had a frequency below 2%.

While somatic copy number alterations play a major role in the pathogenesis of breast tumors[4,5], some germline CNVs are known to be associated with inherited risk of breast cancer. Rare loss of function variants in susceptibility genes such as *BRCA1* and *CHEK2* are associated with a large increase in risk[6]. While the majority of these variants are single nucleotide mutations and short indels, they also include longer deletions and duplications. It has been reported that up to a third of loss of function *BRCA1* variants in some populations may be CNVs [7].

Large-scale genome-wide association studies (GWAS) have established breast cancer associations with common variants at more than 150 loci, mostly in non-coding regions[8-11]. At two of the loci, deletions imputed from the 1000 Genomes reference panel have been identified as likely causal variants. A deletion of the *APOBEC3B* gene-coding region increases breast cancer risk[12] and analysis of the tumours of the germline deletion carriers showed an increase in APOBEC-mediated somatic mutations.[13] A deletion in a regulatory region was identified as a likely causal variant at the 2q35 locus[14,15].

Detecting CNVs from the intensity measurements of genotyping array probes is prone to producing unreliable calls due to the high level of noise. We recently developed a new CNV calling method, CamCNV[16], which focuses on rare CNVs and identifies outlier samples that may have a CNV, based on the intensity distribution across all samples at each probe. We showed that this approach is able to detect CNVs using as few as three probes[16]. Here, we apply this approach to a very large array genotype dataset to search for breast cancer associated CNVs. The analyses are outlined in Figure 1.

Results

Summary of CNVs Detected

After quality control we detected a mean of 2.9 deletions (standard deviation 1.6) and 2.5 duplications (SD 2.0) per sample. Supplementary Data 5 shows the mean length, probe coverage and segment z-scores of called CNVs. Duplications tended to be longer than deletions: for example, deletions called on OncoArray covered a mean of 45 Kilobases (Kb) (SD 106 Kb) over 9.8 probes (SD 17.2), while duplications covered a mean of 109 Kb (SD 202 Kb) over 18.9 probes (SD 36.5). CNV calls observed in multiple samples were concentrated in a small proportion of probes (Supplementary Data 6), with <11% of probes having frequency >0.01% and <2% of probes having frequency >0.5%.

We identified called CNVs which overlapped for at least 90% of their length with rare deletions and duplications (frequency <1%) identified by the 1000 Genomes Project (Supplementary Data 5, Supplementary Figure 1.). Forty-nine percent of OncoArray deletions and 47% of iCOGs deletions matched a 1000 Genomes Project variant while 29% of OncoArray duplications and 20% of iCOGs duplications matched. In total we identified CNVs closely matching 3,273 of the deletion variants published by the 1000 Genomes Project (~9% of total) and 1,255 of their duplication variants (~24% of the total).

CNVs Associated with Overall Risk

Association results were derived for 1,301 regions containing deletions and 992 regions containing duplications. QQ plots are shown in Figure 2a for deletions and 2b for duplications. There was no evidence for inflation in the test statistics for duplications (lambda=0.98; $lambda_{1000}$=1.00) and minimal evidence for deletions (lambda = 1.11; $lambda_{1000}$=1.00).

Seven deletion and two duplication regions were associated with breast cancer risk at p<0.001 (Table 1); of these, deletions within the *BRCA1* region achieved $p< 10^{-6}$. The results for all regions are shown in Supplementary Data 7 and 8 and include statistics on the number of probes covered by the calls. The results for individual probes covered by the regions analysed are in Supplementary Data 9 to 12.

The *BRCA1* locus contains multiple deletions across the gene. The *CHEK2* region (OR 1.94, p=0.0003) covers the whole gene but nearly all the calls correspond to a deletion of exons nine and ten, which was previously observed in 1% of breast cancer cases and 0.4% of controls in Poland [22]. We observed the deletion in 0.9% of Polish cases and 0.5% of controls.

The most significant association (OR=0.69 P=0.00001) for duplications covers a large region on 17p13.3 with multiple long variants overlapping shorter duplications. The OncoArray results by probe show the strongest associations at a series of probes (17: 814529-850542)

in the first intron of *NXN*, with the lowest P-value at 17: 819141 (OR=0.45, P=0.002). The most significant probe position on iCOGs was also in this region (17:836631, OR=0.58, P=0.09) (Figure 3).

Associations with Risk of Breast Cancer Subtypes

We repeated the analyses restricting cases to those with ER-positive, ER-negative and triple negative disease. Deletions and duplications with p-values below 0.001 are shown in Tables 2 and 3 and the results for all regions are in Supplementary Data 13 and 14. An association was observed for *BRCA1* for all subtypes, with the exception of duplications for ER-positive disease. The odds ratio for *BRCA1* deletions was higher for ER-negative disease (OR=27.03; 95% CI, 15.66 to 46.67) than ER-positive (OR =2.81; CI, 1.56 to 5.08; P=8.46E-28 for the difference), while for *CHEK2* the odds ratio was higher for ER-positive disease (OR=2.32;CI,1.56 to 3.44) than ER-negative (OR=1.36; CI,0.66 to 2.82; P=0.11 for the difference), consistent with the known subtype-specific associations for deleterious variants in these genes [23].

In total we observed five deletion and two duplication regions with p-values < 0.001 that did not reach p<0.001 in the overall risk analysis. The strongest novel association for ER-positive was for an intronic deletion in *ITGBL1* (OR = 3.3, P=0.00007, P for difference by ER-status=0.18). For ER-negative disease the strongest novel association was with an intergenic deletion between *ABCC4* (*MRP4*) and *CLDN10* (OR=2.16, P=0.0002, P for difference by ER-status=0.02). Neither of these associations was significant for the other subtype. For triple negative disease, the strongest evidence of association was found for an intergenic duplication between *TMC3* and *MEX3B* (OR=2.39 P= 0.00009) and for two separate deletions upstream of the *DDX18* gene: 2:118258797-118389164 (OR= 6.56, P= 0.00001) and 2:117973154- 118107795 (OR=4.54, P= 0.0008). The association at these two

deletions was driven by the same samples, with 75% of the carriers of the first deletion observed to have the second deletion and normal copy number at the 62kb gap between the deletions.

Associations at Established Common Susceptibility Loci

Three of the most significant associations were observed within regions harbouring known breast cancer susceptibility loci for breast cancer. The most significant (OR=1.42;CI,1.21 to 1.67; P=0.00015) was upstream of *FGFR2* and consistent with a 28 kb deletion in the 1000 Genomes Project data (chr10:123433204-123461492). Three independent risk signals have been previously identified at this region[24,25]. The effect size for the CNV was larger than those previously reported for these common SNPs (largest OR=1.27;CI,1.22 to 1.25). The CNV is in linkage disequilibrium with two of the SNPs identified as likely causally associated variants: rs35054928 (D' = 0.82) and rs2981578 (D' = 0.88). Conditioning on those SNPs reduced somewhat the strength of the association for the deletion (OR =1.30;CI 1.10 to 1.53;P=0.002, Supplementary Data 15).

The third strongest signal (OR=4.9 P=0.00001) in the deletion analysis for overall breast cancer was at 8: 132199447-132252439, 144Kb downstream of *ADCY8*. The strongest GWAS signal in this region lies in an intron of *ADCY8* (lead SNP rs73348588, OR =1.13, P= 8.2e-7)[9]. A 3kb deletion in intron 4 of *KLF12* was associated with ER-negative disease (OR = 2.4, P= 0.0007, P for difference by ER-status=0.01). This is 389kb distant from common SNPs, located between *KLF12* and *KLF5*, associated with ER-negative disease (rs9573140, OR = 0.94, P=3.62e-9) [26]. The *KLF12* and *ADCY8* deletions are not in strong linkage disequilibrium with the corresponding GWAS signals and conditioning on these SNPs did not alter the strengths of the association for the CNVs (Supplementary Data 15) .

Gene Burden tests

We performed gene burden tests based on CNVs that overlapped exons. Analyses were restricted to genes in which at least 24 CNVs were identified, leaving 645 genes with deletions (Supplementary Data 16) and 1596 genes with duplications (Supplementary Data 17). QQ plots are shown in Figures 2c for deletions and 2d for duplications. The lambda for inflation was 1.18 (; $lambda_{1000}$=1.00) for deletions and 1.07 (; $lambda_{1000}$=1.00) for duplications.

For deletions, we found 10 genes with P < 0.01 (Table 4), the most significant being *BRCA1* (OR=7.66, P= 3.72E-18). Four of these 10 genes (*ATM, BRCA1, BRCA2, CHEK2*) are known breast cancer susceptibility genes.[23] Deletions were also observed in two other known susceptibility genes: *PALB2* (23 cases, 9 controls, OR=2.02, P=0.09) and *RAD51C* (21 cases, 9 controls, OR=2.04, P=0.08). The most significant novel association was for *SUPT3H* (OR=0.27, P=0.0004).

For duplications we observed 15 genes with P < 0.01 (Table 5). The most significant association was for *VPS53* (OR = 0.5, P= 0.0009). This gene and *ABR* (OR=0.61 P= 0.008) both lie within the region on 17p which had the strongest association in the regional analysis. These associations were driven by duplications in different samples, with only one duplication in one sample overlapping both genes. Duplications were associated with an increased risk for only four of the 15 genes; the most statistically significant was *RSU1* (OR=3.4, P= 0.004). There was also some evidence of association with risk for duplications in *BRCA1* (OR = 1.75, P =0.01). However, analysis restricted to duplications that included exon 12 of *BRCA1* showed clearer evidence of association (34 carriers, OR = 4.7 P= 0.0001), consistent with one of the more frequent known *BRCA1* duplications that results in a frameshift[27].

The gene burden subtype results are shown in Supplementary Data 18 and 19. The strongest associations were observed for *BRCA1* deletions for ER Negative (OR = 33,

P=5.5E-35) and Triple Negative (OR =49 P=7.1E-34) disease, *CHEK2* deletions for ER positive disease (OR = 2.14 P=0.0001) and *ATM* deletions for ER positive disease (OR=4.85, P=0.0001). No additional genes significant at P<0.0001 were found.

Direction of effect tests

In the gene burden and individual probe analyses we observed a directional effect, whereby the strongest associations for deletions tended to increase risk and those for duplications tended to be protective. To test whether these associations deviated from what would be expected by chance, we computed ranked summed z-score tests and evaluated the significance of the maximum test statistic by permutation. Results are summarised in Table 6. The statistic for deletion regions was more significant than any of the permuted statistics (P=0.04) but was reduced to P=0.12 after removing the known genes *BRCA1* and *CHEK2*. The significance of the gene burden test for deletions also was reduced from P=0.04 to P=0.2 when the known genes were removed. The statistic for the duplication regions was lower than any of the 50 permutations (P=0.04). The gene burden analysis for duplications was not significant.

Discussion

We used the largest available breast cancer case-control dataset, comprising more than 86,000 cases and 76,000 controls with array genotyping, to test for associations with rare CNVs. Using the intensities from genotyping arrays to detect CNVs is not ideal due to a high level of noise and uncertainty in the calling, particularly for duplications. However, in tests of known CNVs and replication of calls between duplicate samples, the CamCNV method shows reasonable levels of sensitivity and specificity[16]. The main focus of this analysis was low frequency CNVs (<1% frequency) since higher frequency CNVs can generally be studied through imputation to a reference panel. In the 0.05%-1% frequency range, we could

406 detect ~20% of the CNVs identified by the 1000 Genomes project. For some loci we only
407 had evidence from one array because the probes do not exist to detect the variants on the
408 other array. Thus, while this array-based approach provides power to evaluate the CNVs
409 that can be assayed, much denser arrays or direct sequencing would be required to provide
410 a complete evaluation of the contribution of CNVs.

411 In support of the reliability of the method, we detected evidence for both deletions and
412 duplications in *BRCA1*, which was stronger for ER-negative disease, and for deletions in
413 *CHEK2* , which were stronger for ER-positive disease. The latter appeared to be driven by a
414 single founder deletion in East European populations. Weaker evidence of association was
415 also observed for deletions in other susceptibility genes (*BRCA2, ATM, PALB2, RAD51C*);
416 the ORs were consistent with those seen for deleterious SNVs and indels. [23] In total, around
417 0.5% of cases in our analysis had a deletion in one of the known susceptibility genes with
418 the proportion rising to ~1% for cases diagnosed under 50 years of age. The majority of
419 coding deletions are expected to affect only part of the gene, with one study observing that
420 a quarter covered only a single exon.[28] To detect all of these using array data would require
421 at least three probes per exon. The OncoArray has this level of coverage for a few genes,
422 including *BRCA1* and *BRCA2,* but the coverage is lower for most genes and many coding
423 CNVs will have been missed.

424 A key issue is the appropriate level of statistical significance to apply to these analyses. For
425 the gene burden tests, $P<2.5x10^{-6}$, as used in exome-sequencing, seems an appropriate
426 level. It is less clear what is appropriate for non-coding variants. A level of $P<5x10^{-8}$ has
427 become standard in GWAS and has been shown to lead to acceptable replication, but this
428 seems over-conservative for CNVs, which are more likely to be deleterious. Consistent with
429 this, for at least two of the ~200 common susceptibility loci, the likely causal variant is a
430 CNV, a higher fraction than expected given the relative frequencies of CNVs and SNPs.
431 Based on frequency analysis of whole-genome sequence data Abel et al. [1] estimated that
432 rare CNVs are >800 times more likely to be deleterious than rare SNVs and >300 times

more likely than rare indels. On the other hand, the significance level for non-coding CNVs should logically be more stringent than for the gene burden tests. Taken together, a significance level of ~10^{-6} seems appropriate, while associations at P<0.001 may be worth following up in future studies. In our analyses only the association at *BRCA1* (both in the overall and gene burden tests) passes the higher threshold. We also calculated Bayesian False Discovery Probabilities (BFDPs)[21] (Supplementary Data 20 and 21) for our associations using prior probabilities of 0.001 for regions and 0.002 for genes. Outside the known genes none of the BDFPs gave a probability below 10%, with the lowest BFDP of 0.11 for the deletion in the *FGFR2* locus. For a CNV observed with a frequency of 0.1% (n=91 samples in the OncoArray dataset) we had 40% power to detect an association with an odds ratio of 2 but only 1.5% power to detect an association with an odds ratio of 1.5. An OR of 2, comparable to that seen for deleterious variants in *ATM* and *CHEK2*, may be plausible for rare coding CNVs or non-coding CNVs that have a significant effect on gene expression. Larger sample sizes will clearly be required to evaluate rare CNVs with more modest associations.

In addition to the *BRCA1* and *CHEK2* loci, we found associations in three known susceptibility regions identified through GWAS, harbouring *FGFR2*, *ADCY8* and *KLF12*. In each case, the variants are rarer than the established associated variants, but confer higher risks. The *ADCY8* and *KLF12* deletions are not in linkage disequilibrium with the associated SNPs. The *FGFR2* deletion is in linkage disequilibrium with two of the likely causal common SNPs although there was still evidence of association with the deletion, albeit weaker, after conditional analysis. *In-silico* and functional analysis clearly demonstrate that FGFR2 is the target of the previously established variants[24,25]; it will be interesting to establish if the same is true for the CNV.

Excluding loci in known susceptibility regions, the strongest evidence of association was for a 12kb deletion (13:102121830-102133956) in the second intron of *ITGBL1* (OR = 3.3, P=0.00007 in the ER-positive analysis). This deletes a promoter flanking region (Ensembl

ID: ENSR00001563823) and CTCF binding site (Ensembl ID: ENSR00001062398) active in mammary epithelial cells. There is experimental evidence that *ITGBL1* expression, mediated by the RUNX2 transcription factor, enables breast cancer cells to form bone metastases[29].

In the gene burden analysis, the strongest novel association was for deletions within *SUPT3H*, which were associated with a reduced risk. *SUPT3H* encodes human SPT3, a component of the STAGA complex which acts as a co-activator of the MYC oncoprotein[30]. *SUPTH* is located within the first intron of the *RUNX2* transcription factor and the syntenic relationship between the two genes is highly evolutionarily conserved [31]. *RUNX2* has a role in mammary gland development and high *RUNX2* expression is found in ER-negative tumours.[32] The *PCDHGB2* association appears to be due to a single variant (5:140739812-140740918) that deletes the first exon but as this gene is part of the protocadherin gamma gene cluster it is also possible that the deletion may be having an effect on one of the five genes that overlap *PCDHGB2*. It also deletes a promoter active in mammary epithelial cells (ENSR00001342785). The next strongest signals were for *MEAK7* (OR=2.19 P= 0.001), a gene implicated in a mTOR signalling pathway[33], and *MAD1L1* ((OR=2.00 P=0.005), a component of the mitotic spindle-assembly that has been suggested as a possible tumour suppressor[34].

After *BRCA1*, the most significant association for ER Negative disease in the gene burden analysis was for *CYP2C18* which overlaps *CYP2C19* (ER-negative OR=2.6, P=0.002; triple-negative OR=4.4, P=0.0002). A previous analysis of CNVs and breast cancer in the Finnish population identified a founder mutation reaching an overall frequency of ~ 3% and reported a possible association at this locus for triple negative (OR 2.8, p= 0.02) and ER-negative breast cancer (OR =2.2 p=0.048).[35]

The results from duplications are harder to interpret as there are often longer duplications overlapping whole genes with shorter variants covering some part of their length. For the gene burden analysis there was little evidence of strong associations. In the regional analysis, the two strongest associations cover multiple genes. The strongest evidence of

association (OR=0.69 P=1.1E-05) was for a 1.5 Mb region at the start of chromosome 17 (17:13905- 1559829). The probe-specific and gene burden results highlighted some stronger signals within this region, for example within *NXN* and *VPS53*, but the direction of effect was consistent across the region with 80% of the OncoArray probes having an odds ratio of 0.75 or lower (Figure 3). This locus has established associations with prostate and colorectal cancer. Interestingly a possible association with ER-positive breast cancer survival was detected for a rare SNP in the first intron of NXN, rs118021774 (HR=1.83, P=3.8E-06)[36]. The detected duplications are not in LD with this SNP. For the 0.4Mb duplication region on chromosome 21 (OR= 2.23 P=0.0001) the probe-specific results from OncoArray highlighted that the signal is specific to a shorter intergenic region (21:33421860- 33459975) between *HUNK* and *LINC00159*.

We observed some evidence of an aggregate directional effect, both for genes and non-genic regions, such that the deletions in aggregate were associated with increased risk. There was also some suggestion that duplications, in aggregate, were associated with a reduced risk. These results suggest that additional associations are present that could be established with a larger dataset. A new GWAS, Confluence (https://dceg.cancer.gov/research/cancer-types/breast-cancer/confluence-project), aims to double the available sample size for breast cancer. This GWAS includes probes specifically designed to assay some of the most significant CNVs observed in this study (those significant at P<0.001), and the sample size should be sufficient to confirm or refute these associations.

Methods

Subjects

Data were derived from blood samples from study participants in 66 studies participating in the Breast Cancer Association Consortium (BCAC) and genotyped as part of the OncoArray[9,17] and iCOGS[8] collaborations (Supplementary Data 1). Studies included population-based and hospital-based case-control studies, and case-control studies nested within prospective cohorts; we only included data from studies that provided both cases and controls. Phenotype data were based on version 12 of the BCAC database. Cases were diagnosed with either invasive breast cancer or carcinoma-in-situ. Oestrogen receptor (ER) status was determined from medical records or tissue microarray evaluation, where available. Analyses were restricted to participants of European ancestry, as defined by ancestry informative principal components[8,9]. Where samples were genotyped on both arrays, we excluded the iCOGS sample as the OncoArray has better genome-wide coverage. After sample quality control (see below), data on 36,980 cases and 34,706 controls with iCOGS genotyping, and 49,808 cases and 41,416 controls with OncoArray genotyping, were available for analysis (Supplementary Data 2).

Arrays

The Illumina iCOGs genotyping array[8] includes 211,155 probes for SNVs and short insertions/deletions. Most variants were selected because of previous association in case-control studies for breast prostate and ovarian cancers, or for dense mapping of regions harbouring an association. The OncoArray includes 533,631 probes, of which approximately half were selected from the Illumina HumanCore backbone, a set of SNPs designed to tag most common variants. The remainder were selected on the basis of evidence of previous association with breast, prostate, ovarian, lung or colorectal cancer risk. Approximately 32,000 variants on the OncoArray were selected to provide dense coverage of associated

loci and known genes. The remainder were mostly selected from lists of common variants ranked by p-value, with a small number from lists of candidate variants.

CNV Calling

CNVs were called using the CamCNV pipeline as previously described[16]. In brief, the log R (LRR) intensity measurements and B allele frequency (BAF) for each sample at each probe were exported from Illumina's Genome Studio software. A principal component adjustment (PCA) was applied to the LRR, grouped by study, to remove noise and batch effects. After removing noisy probes and those in regions with known common CNVs, the LRRs for each probe were converted to z-scores using the mean and standard deviation from all BCAC samples. Circular binary segmentation was applied to the z-scores ordered by probe position for each sample using the DNACopy package.[18] This produces a list of segments for each sample by chromosome where the z-score of consecutive probes changes by more than two standard deviations. Segments with a mean probe z-score between -3.7 and -14 were called as deletions and segments with a mean z-score between +2 and +10 as duplications. We restricted the calls to segments covering a minimum of three and a maximum of 200 probes.

As per the CamCNV pipeline, we then excluded deletions with inconsistent B Allele Frequency and CNVs with a shift in LRR at the sample level that was outside the expected range. The additional CNV exclusions are summarised in Supplementary Data 3. To exclude regions with a high level of noise we also excluded CNVs falling within 1Mb of telomeres and centromeres and a number of immune loci such as the T-cell receptor genes where somatic mutations in blood are often observed [19].

Sample Quality Control

Standard sample quality control exclusions were performed, as previously described for the SNP genotype analyses[8,9]. These include exclusions for excess heterozygosity, ancestry outliers, mismatches with other genotyping, and close relatives. A stricter sample call rate of

>99% was used for the CNV analysis, compared to >95% used in the genotype analyses. We also excluded any participants for whom a DNA sample was not collected from blood and any that had been whole genome amplified.

In addition, we used two metrics to exclude noisy samples liable to produce an excess of unreliable CNV calls. First, we calculated a derivative log ratio spread (DLRS) figure for each sample as the standard deviation of the differences between LRR for probes ordered by genomic position, divided by the square root of two. This measures the variance in the LRR from each probe to the next averaged over the whole genome and thus is insensitive to large fluctuations such as might be expected between different chromosomes in the same sample. An ideal sample would have a small DLRS as the only variance would come from a small number of genuine CNVs. We calculated the DLRS using the dLRs function in R package ADM3 (https://CRAN.R-project.org/package=ADM3) before and after the PCA. At both stages we excluded samples with a DLRS more than 3.5 standard deviations above the mean DLRS for that study.

Second, we counted the number of short segments (between three and 200 probes) output by DNACopy for each sample. We observed that the distribution of segment counts was skewed to the right with an excess of samples with a large number of segments. We calculated a cut-off for the maximum number of segments using the following formula where x is the segment count for each sample (based on the rationale that the distribution of the true number of segments should be approximately Poisson):

y=2*sqrt(x)

cut-off = median(y)+3.5

The sample exclusions resulting from these QC steps are summarised in Supplementary Data 4.

Association Tests

All analyses were carried out separately for deletions and duplications, since different types of CNV at the same locus do not necessarily have the same effect on risk. As we were only assessing rare CNVs, we treated all carriers as heterozygotes and did not attempt to identify rare homozygotes.

To account for overlapping CNVs and imprecision in the breakpoints, we assigned individual CNVs to regions. To identify the regions, we moved sequentially along each chromosome, identifying the start as an Oncoarray probe position where deletions were observed in at least five samples, and then the end position as the probe position before the first probe where deletions were observed in fewer than five samples. Regions within five probes of each other were then merged together. The process was repeated for duplications. Regions were also merged such that the major susceptibility genes (*BRCA1, BRCA2, CHEK2*) were included within a single region. We then assigned individual CNVs to regions where at least 90% of the CNV's length fell within the region. For iCOGS, which generally has less dense probe coverage, we first assigned CNVs to the OncoArray regions where they showed > 90% overlap. We then assigned any remaining CNVs to regions defined using the iCOGS probes, using the same procedure. Using this approach, 3,306 deletion regions were identified from OncoArray data, 812 of which were also observed using iCOGS data, and 541 regions identified using iCOGS alone. For duplications there were 2,203 OncoArray regions, with 854 also observed using iCOGS data, and 483 iCOGS specific regions.

Associations were evaluated for each region using logistic regression, with breast cancer status as the outcome, and the presence of a CNV in the region (0 or 1) as a covariate to derive a log odds ratio per deletion/duplication. Statistical significance was evaluated using a likelihood ratio test (based on the above model and one excluding CNV as a covariate). The logistic regression analyses were conducted using in-house software (https://ccge.medschl.cam.ac.uk/software/mlogit/). Study and ten ancestry informative principal components, defined separately for each array, were also included as covariates.

611 The analyses were conducted separately for the iCOGS and OncoArray and then combined
612 in a standard fixed effect meta-analysis using the METAL software (after first deriving the
613 standard error of the log-odds ratio from the likelihood test statistic)[20]. To avoid regions with
614 too few observations, we excluded regions with fewer than 24 deletions or duplications
615 (~0.015% of samples). Associations significant at P<0.001 were considered noteworthy.

616 To detect more precisely the location of association signals, we also generated results for
617 each probe. We created a vector of pseudo-genotypes for each probe with samples, such
618 that a deletion covering that probe was coded as 1 and all other samples were coded as 0.
619 We generated a similar set of genotypes for duplications. The results were analysed using
620 logistic regression, as above.

621 To test for association between CNVs affecting the coding sequence of genes, in aggregate,
622 and breast cancer risk, we identified samples with a deletion or duplication overlapping the
623 exons of each gene. Exon positions were downloaded from the UCSC Genome Browser
624 hg19 knownGene table. We used logistic regression to generate a log odds ratio (OR) for
625 carriers of coding variants covering each gene, adjusted for study, as above. We generated
626 results for each array and then for carriers combined across both arrays. For the combined
627 analyses we treated studies with samples on both arrays as separate studies.

628 To calculate Bayesian False Discovery Probabilities (BFDPs) we assumed a log-normally
629 distributed prior effect size as described by Wakefield[21]. The prior log(OR) was determined
630 by assuming a 95% probability that the OR was less than some bound K, where K=3 for the
631 regional and gene-based analysis, except for *BRCA1* and *BRCA2* where K=20 was
632 assumed. The prior probability of association was assumed to be 0.001 for the regional
633 analysis, 0.99 for *BRCA1, BRCA2, ATM* and *CHEK2* and 0.002 for other genes. For the
634 gene-based analysis only positive associations were considered as the prior evidence for all
635 genes was in favour of PTVs being positively associated with risk.

To determine whether there was a tendency for CNVs to be associated with an excess, or deficit, of risk across genes or regions, we computed signed z-scores as the square root of the chi-squared statistic for each gene, multiplied by ±1 depending on whether the effect estimate was positive or negative. These were ranked and normalised summed z-scores, based on the r most significant associations, were derived. The overall test statistic was the maximum summed z-score over all possible values of *r*:

Equation 1:

$$U = \max_{r \leq n} \frac{1}{\sqrt{r}} \sum_{j=1}^{r} z_j$$

Where *n* is the total number of genes/regions being tested. The significance of U was then determined by permutation, randomly permuting case-control labels within study 50 times.

Ethical Approval

All participating studies were approved by their appropriate ethics review board and all subjects provided written informed consent.

Tables

Table 1. CNVs associated with overall risk

Type	Locus	Chr.	Start (Build37)	End	Total Carriers	Odds Ratio	Lower CI	Upper CI	Direction (OncoArray, iCOGs)	P-value
Deletion	BRCA1	17	41188342	41363651	195	6.27	4.02	9.79	++	6.32E-16
Deletion	Intergenic_FGFR2_ATE1	10	123435817	123461066	630	1.42	1.21	1.67	++	1.42E-05
Deletion	Intergenic_ADCY8_EFR3A	8	132199447	132250643	42	4.88	2.24	10.61	++	6.36E-05
Deletion	KLHL1	13	70652321	71029916	1761	0.85	0.77	0.92	--	2.31E-04
Deletion	CHEK2	22	29083731	29123846	141	1.94	1.35	2.79	++	3.26E-04
Deletion	SUPT3H	6	44908728	45244478	32	0.23	0.1	0.52	-?	4.25E-04
Deletion	Intergenic_GALNT1_C18orf21	18	33350917	33359197	123	1.92	1.32	2.78	?+	6.24E-04
Duplication	17p13.3_VPS53;NXN	17	13905	1559829	577	0.69	0.59	0.82	--	1.08E-05
Duplication	21q22.11_HUNK_LINC00159	21	33410933	33863246	102	2.23	1.47	3.38	++	1.48E-04

653 Table 2. Subtype associations for deletions

Subtype	Locus	Chr.	Start (Build37)	End	Total Carriers	Odds Ratio	Lower CI	Upper CI	Direction (OncoArray, iCOGs)	P-value
ER Positive	Intergenic_FGFR2_ATE1	10	123435817	123461066	478	1.52	1.27	1.81	++	5.04E-06
ER Positive	CHEK2	22	29091788	29102967	79	2.36	1.56	3.44	++	3.03E-05
ER Positive	ITGBL1 intronic	13	102122905	102133221	58	3.29	1.83	5.92	+?	7.29E-05
ER Positive	Intergenic_ADCY8_EFR3A	8	132199447	132250643	31	4.95	2.20	11.30	++	1.07E-04
ER Positive	BRCA1	17	41188342	41363651	57	2.81	1.55	5.08	++	6.41E-04
ER Negative	BRCA1	17	41188342	41363651	112	27.03	15.66	46.67	++	2.62E-32
ER Negative	Intergenic:ABCC4_CLDN10	13	95991263	96004144	134	2.16	1.43	3.26	+?	2.46E-04
ER Negative	KLF12 intronic	13	74356683	74357984	93	2.39	1.49	3.82	+?	2.89E-04
ER Negative	Intergenic_DPP10_DDX18	2	118258797	118389164	44	3.34	1.64	6.80	++	8.84E-04
Triple Neg.	BRCA1	17	41188342	41363651	72	40.55	21.70	75.76	++	3.64E-31
Triple Neg.	Intergenic_DPP10_DDX18	2	118258797	118389164	40	6.56	2.83	15.18	++	1.13E-05
Triple Neg.	Intergenic_DPP10_DDX18	2	117973154	118107795	48	4.54	1.88	10.97	++	7.92E-04

654

655

Table 3. Subtype associations for duplications

Subtype	Locus	Chr.	Start (Build37)	End	Total Carriers	Odds Ratio	Lower CI	Upper CI	Direction (OncoArray, iCOGs)	P-value
ER Positive	17p13.3_VPS53;NXN	17	13905	1559829	454	0.67	0.55	0.81	--	4.44E-05
ER Positive	21q22.11_HUNK_LINC00159	21	33410933	33863246	77	2.55	1.6	4.06	++	7.88E-05
ER Positive	15q13	15	28440287	32797352	1250	0.83	0.75	0.93	--	6.95E-04
ER Negative	BRCA1	17	41188342	41363651	42	5.93	2.31	15.19	+-	2.09E-04
Triple Neg.	BRCA1	17	41188342	41363651	35	10.80	3.33	35.02	+-	7.29E-05
Triple Neg.	Intergenic_TMC3_MEX3B	15	81960409	82104822	231	2.39	1.55	3.71	++	9.25E-05

658 Table 4. Gene burden results for deletions

Gene	Cases	Controls	Odds Ratio	Lower CI	Upper CI	P-value
BRCA1	171	22	7.66	4.84	12.13	3.72E-18
CHEK2	103	48	1.83	1.29	2.61	7.66E-04
SUPT3H	10	25	0.27	0.13	0.59	9.24E-04
PCDHGB2	25	3	7.03	2.10	23.52	1.55E-03
MEAK7	59	24	2.19	1.34	3.58	1.66E-03
ATM	35	8	3.43	1.56	7.51	2.11E-03
MAD1L1	54	25	2.00	1.23	3.26	5.53E-03
NPHP1	477	351	1.22	1.06	1.41	6.13E-03
ZNF320	29	7	3.28	1.39	7.73	6.63E-03
BRCA2	65	33	1.81	1.17	2.81	7.82E-03

659

660 Table 5. Gene burden results for duplications

Gene	Cases	Controls	Odds Ratio	Lower CI	Upper CI	P-value
VPS53	40	65	0.50	0.33	0.75	9.46E-04
ATP12A	48	66	0.57	0.39	0.84	3.97E-03
USP18	12	23	0.34	0.17	0.71	4.16E-03
RPS6KA2	10	22	0.32	0.14	0.70	4.20E-03
RSU1	21	8	3.40	1.47	7.84	4.23E-03
AC008132.1	7	17	0.26	0.10	0.66	4.49E-03
PNPLA4	479	346	1.23	1.06	1.42	5.30E-03
NLGN4X	291	320	0.79	0.67	0.93	5.55E-03
ZNF439	31	9	2.98	1.37	6.45	5.72E-03
TRIM6	5	19	0.24	0.09	0.67	6.34E-03
RP11.363G10.2	13	27	0.39	0.20	0.78	7.39E-03
USP31	7	18	0.30	0.12	0.73	8.02E-03

TRDN	15	29	0.42	0.22	0.80	8.26E-03
ABR	51	69	0.61	0.42	0.88	8.88E-03
DNAJC15	109	67	1.52	1.11	2.09	9.29E-03

661

662 Table 6. Direction of Effect Results

Category	**Analysis ustat**	**Max.ustat of 50 simulations**	**Min. ustat of 50 simulations**	**Simulations with larger/smaller ustat**	**P-value**
Deletion regions	9.48	6.96	-7.81	0	0.04
Deletion regions minus known[1]	5.9	6.96	-7.81	2	0.12
Duplication regions	-9.2	6.54	-6.99	0	0.04
Gene Deletions	9.18	5.64	-9.67	0	0.04
Gene Deletions minus known[2]	5.01	5.64	-9.67	4	0.20
Gene Duplications	-4.33	5.96	-11.26	33	1.29

663 1. BRCA1, CHEK2 regions excluded; 2. BRCA1, CHEK2, ATM, BRCA2 removed.

664

Figures

Figure 1: Flow diagram showing the calling and analysis pipeline.

Figure 2. QQ plots for association of deletion regions (a), duplication regions (b), gene burden analysis for deletions (c) and gene burden for duplications (d).

Figure 3. Plot of log odds ratios for probes within the 17p13.3 duplication region showing genes and 1000 Genomes CNVs from Ensembl.

References

1. Abel, H.J. *et al.* Mapping and characterization of structural variation in 17,795 human genomes. *Nature* (2020).
2. Collins, R.L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444-451 (2020).
3. Sudmant, P.H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**(2015).
4. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).
5. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128 (2020).
6. Easton, D.F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *The New England journal of medicine* **372**, 2243-2257 (2015).
7. Ewald, I.P. *et al.* Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet Mol Biol* **32**, 437-46 (2009).
8. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**(2013).
9. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92 (2017).
10. Milne, R.L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nature Genetics* **49**, 1767 (2017).
11. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics* **52**, 572-581 (2020).
12. Long, J. *et al.* A common deletion in the APOBEC3 genes and breast cancer risk. *J Natl Cancer Inst* **105**(2013).
13. Nik-Zainal, S. *et al.* Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet* **46**, 487-91 (2014).
14. Wyszynski, A. *et al.* An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Hum Mol Genet* (2016).
15. Baxter, J. Functional annotation of the 2q35 breast cancer risk locus implicates a structural variant in influencing activity of a long-range enhancer element. *Am J Hum Genet* **108,** 1190-1203 (2021).

16. Dennis, J., Walker, L., Tyrer, J., Michailidou, K. & Easton, D.F. Detecting rare copy number variants from Illumina genotyping arrays with the CamCNV pipeline: Segmentation of z-scores improves detection and reliability. *Genet Epidemiol* (2020).
17. Amos, C.I. *et al.* The OncoArray Consortium: A network for understanding the genetic architecture of common cancers. *Cancer Epidemiol Biomarkers Prev* **26**(2017).
18. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23**, 657-663 (2007).
19. Schwienbacher, C. *et al.* Copy number variation and association over T-cell receptor genes--influence of DNA source. *Immunogenetics* **62**, 561-7 (2010).
20. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
21. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet* **81**, 208-27 (2007).
22. Cybulski, C. *et al.* A deletion in CHEK2 of 5,395 bp predisposes to breast cancer in Poland. *Breast Cancer Res Treat* **102**, 119-22 (2007).
23. Dorling, L. *et al.* Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med* **384**, 428-439 (2021).
24. Meyer, K.B. *et al.* Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am J Hum Genet* **93**, 1046-60 (2013).
25. Fachal, L. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics* **52**, 56-73 (2020).
26. Milne, R.L. *et al.* Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* **49**, 1767-1778 (2017).
27. The Exon 13 Duplication in the BRCA1 Gene Is a Founder Mutation Present in Geographically Diverse Populations. *Am J Hum Genet* **67**, 207-12 (2000).
28. Truty, R. *et al.* Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genetics in Medicine* (2018).
29. Li, X.Q. *et al.* ITGBL1 Is a Runx2 Transcriptional Target and Promotes Breast Cancer Bone Metastasis by Activating the TGFβ Signaling Pathway. *Cancer Res* **75**, 3302-13 (2015).
30. Liu, X., Vorontchikhina, M., Wang, Y.-L., Faiola, F. & Martinez, E. STAGA Recruits Mediator to the MYC Oncoprotein To Stimulate Transcription and Cell Proliferation. *Molecular and Cellular Biology* **28**, 108-121 (2008).
31. Barutcu, A.R. *et al.* The bone-specific Runx2-P1 promoter displays conserved three-dimensional chromatin structure with the syntenic Supt3h promoter. *Nucleic Acids Res* **42**, 10360-72 (2014).
32. Ferrari, N., McDonald, L., Morris, J.S., Cameron, E.R. & Blyth, K. RUNX2 in mammary gland development and breast cancer. *J Cell Physiol* **228**, 1137-42 (2013).
33. Nguyen, J.T. *et al.* Mammalian EAK-7 activates alternative mTOR signaling to regulate cell proliferation and migration. *Science Advances* **4**, eaao5838 (2018).
34. Rajaram, M. *et al.* Two Distinct Categories of Focal Deletions in Cancer Genomes. *PLOS ONE* **8**, e66264 (2013).
35. Tervasmaki, A., Winqvist, R., Jukkola-Vuorinen, A. & Pylkas, K. Recurrent CYP2C19 deletion allele is associated with triple-negative breast cancer. *BMC Cancer* **14**, 902 (2014).
36. Escala-Garcia, M. *et al.* Genome-wide association study of germline variants and breast cancer-specific mortality. *British Journal of Cancer* **120**, 647-657 (2019).

Data Availability

Full summary statistics for the regions and probes analysed are available in the Supplementary Data. This includes the source data used to produce Figures 2 and 3. The majority of the OncoArray dataset analyzed in this study is available in the dbGap repository,

Study ID: phs001265.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1). The iCOGS dataset and complete OncoArray dataset cannot be made publicly available due to restraints imposed by the ethics committees of individual studies; requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of BCAC (http://bcac.ccge.medschl.cam.ac.uk/).

Code Availability

Code for the CamCNV calling algorithm and a test dataset with OncoArray genotyping is available at the project home page (https://github.com/jgd29/CamCNV).

Competing Interests

The authors declare that there are no conflicts of interests.

Contributions

Data management: M.K.B., Q.Wan., R.Ke. CNV detection: J.D., L.W. Statistical/bioinformatic analysis: J.D., J.T., K.M., L.Dorl. Conceived the OncoArray and obtained financial support: J.S., P.K. and D.F.E. Provided DNA samples and/or phenotypic data: T.A., I.A., H.A., N.A., V.A., K.A., L.B., M.W.B., S.Beh., J.B., M.B., N.B., S.Boj., H.B., J.E.C., J.C-C., G.C., C.L.C., J.M.C., F.C., A.C., S.C., K.C., P.D., T.D., L.Doss., A.E., M.E., D.G.E., P.F., J.F., O.F., H.F., L.F., M.G., M.G-D., M.G-G., G.G., A.G., P.G., E.H., C.H., P.H., A.Hol., R.H., J.H., A.How.., A.Jag., A.Jak., E.J., N.J., M.J., A.Jung, R.Ka., E.K., C.K., Y.K., V-M.K., S.K., P.K., V.N.K., K.K-S., A.K., J.L., D.L., N.L., M.L., A.L., A.M., S.Man., S.Mar., D.M., R.L.M., T.M., R.A.M., H.N., J.O., H.O., T.P., C.P., P.P.,P.D.P.P, D.P-K., K.P., G.R., E.Sal., D.S., E.J.S., M.S., R.K.S., R.S.,J.S., A.S., P.S., M.C.S., A.J.S., R.M.T., J.A.T., L.R.T., M.B.T., I.T., M.A.T., T.T., C.V., C.W., R.W., A.W., X.Y., W.Z., A.Z., A.D., D.F.E., ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, CTS Consortium. Directed the project: D.F.E.

All authors read and approved the final version of the manuscript.

Acknowledgements

Joe Dennis is supported by the CanRisk Cancer Research UK programme grant: PPRPGM-Nov20\100002 and by the Confluence project which is funded with intramural funds from the National Cancer Institute Intramural Research Program, National Institutes of Health. Logan Walker is supported by a Rutherford Discovery Fellowship (Royal Society of New Zealand).

We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. The COGS study would not have been possible without the contributions of the following Andrew Berchuck (OCAC), Rosalind A. Eeles, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Antonis Antoniou, Lesley McGuffog and Ken Offit (CIMBA), Andrew Lee, and Ed Dicks, Craig Luccarini and the staff of the Centre for Genetic Epidemiology Laboratory, the staff of the CNIO genotyping unit, Jacques Simard and Daniel C. Tessier, Francois Bacot, Daniel Vincent, Sylvie LaBoissière and Frederic Robidoux and the staff of the McGill University and Génome Québec Innovation Centre, Sune F. Nielsen, Borge G. Nordestgaard, and the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility. ABCFS thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCS thanks the Blood bank Sanquin, The Netherlands. ABCTB Investigators: Christine Clarke, Deborah Marsh, Rodney Scott, Robert Baxter, Desmond Yip, Jane Carpenter, Alison Davis, Nirmala Pathmanathan, Peter Simpson, J. Dinny Graham, Mythily Sachchithananthan. Samples are made available to researchers on a non-exclusive basis. BBCS thanks Eileen

Williams, Elaine Ryder-Mills, Kara Sargus. BCEES thanks Allyson Thomson, Christobel Saunders, Terry Slevin, BreastScreen Western Australia, Elizabeth Wylie, Rachel Lloyd. The BCINIS study would not have been possible without the contributions of Dr. K. Landsman, Dr. N. Gronich, Dr. A. Flugelman, Dr. W. Saliba, Dr. F. Lejbkowicz, Dr. E. Liani, Dr. I. Cohen, Dr. S. Kalet, Dr. V. Friedman, Dr. O. Barnet of the NICCC in Haifa, and all the contributing family medicine, surgery, pathology and oncology teams in all medical institutes in Northern Israel. BIGGS thanks Niall McInerney, Gabrielle Colleran, Andrew Rowan, Angela Jones. The BREOGAN study would not have been possible without the contributions of the following: Angel Carracedo, Victor Muñoz Garzón, Alejandro Novo Domínguez, Maria Elena Martinez, Sara Miranda Ponte, Carmen Redondo Marey, Maite Peña Fernández, Manuel Enguix Castelo, Maria Torres, Manuel Calaza (BREOGAN), José Antúnez, Máximo Fraga and the staff of the Department of Pathology and Biobank of the University Hospital Complex of Santiago-CHUS, Instituto de Investigación Sanitaria de Santiago, IDIS, Xerencia de Xestion Integrada de Santiago-SERGAS; Joaquín González-Carreró and the staff of the Department of Pathology and Biobank of University Hospital Complex of Vigo, Instituto de Investigacion Biomedica Galicia Sur, SERGAS, Vigo, Spain. CBCS thanks study participants, co-investigators, collaborators and staff of the Canadian Breast Cancer Study, and project coordinators Agnes Lai and Celine Morissette. CCGP thanks Styliani Apostolaki, Anna Margiolaki, Georgios Nintos, Maria Perraki, Georgia Saloustrou, Georgia Sevastaki, Konstantinos Pompodakis. CGPS thanks staff and participants of the Copenhagen General Population Study. For the excellent technical assistance: Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank, Dorthe Kjeldgård Hansen. The Danish Cancer Biobank is acknowledged for providing infrastructure for the collection of blood samples for the cases. CNIO-BCS thanks Guillermo Pita, Charo Alonso, Nuria Álvarez, Pilar Zamora, Primitiva Menendez, the Human Genotyping-CEGEN Unit (CNIO). Investigators from the CPS-II cohort thank the participants and Study Management Group for their invaluable contributions to this research. They also acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, as well as cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program. The authors would like to thank the California Teachers Study Steering Committee that is responsible for the formation and maintenance of the Study within which this research was conducted. A full list of California Teachers Study team members is available at https://www.calteachersstudy.org/team. ESTHER thanks Hartwig Ziegler, Sonja Wolf, Volker Hermann, Christa Stegmaier, Katja Butterbach. FHRISK thanks NIHR for funding and the Manchester NIHR Biomedical Research Centre (IS-BRC-1215-20007). GC-HBOC thanks Stefanie Engert, Heide Hellebrand, Sandra Kröber and LIFE - Leipzig Research Centre for Civilization Diseases (Markus Loeffler, Joachim Thiery, Matthias Nüchter, Ronny Baber). The GENICA Network: Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany [Hiltrud Brauch, RH, Wing-Yee Lo], German Cancer Consortium (DKTK) and German Cancer Reseach Center (DKFZ), Partner Site Tübingen, 72074 Tübingen, Germany [Hiltrud Brauch], gefördert durch die Deutsche Forschungsgemeinschaft (DFG) im Rahmen der Exzellenzstrategie des Bundes und der Länder – EXC 2180 – 390900677 [Hiltrud Brauch], Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany [YDK, Christian Baisch], Institute of Pathology, University of Bonn, Germany [Hans-Peter Fischer], Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany [Ute Hamann], Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany [Thomas Brüning, Beate Pesch, Sylvia Rabstein, Anne Lotz]; and Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany [Volker Harth]. HEBCS thanks Johanna Kiiski, Carl Blomqvist, Kristiina Aittomäki, Kirsimari Aaltonen, Karl von Smitten, Irja Erkkilä. HMBCS thanks Peter Hillemanns, Hans Christiansen and Johann H. Karstens. HUBCS thanks Shamil Gantsev. ICICLE thanks Kelly Kohut, Michele Caneppele, Maria Troy. KARMA and

SASBAC thank the Swedish Medical Research Counsel. KBCP thanks Eija Myöhänen. kConFab/AOCS wish to thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab. LMBC thanks Gilian Peuteman, Thomas Van Brussel, EvyVanderheyden and Kathleen Corthouts. MABCS thanks Milena Jakimovska (RCGEB "Georgi D. Efremov"), Snezhana Smichkoska, Emilija Lazarova (University Clinic of Radiotherapy and Oncology), Mitko Karadjozov (Adzibadem-Sistina Hospital), Andrej Arsovski and Liljana Stojanovska (Re-Medika Hospital) for their contributions and commitment to this study. MARIE thanks Petra Seibold, Dieter Flesch-Janys, Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers and Stefan Nickels. MBCSG (Milan Breast Cancer Study Group): Paolo Radice, Bernard Peissel, Jacopo Azzollini, Erica Rosina, Daniela Zaffaroni, Bernardo Bonanni, Irene Feroce, Mariarosaria Calvello, Aliana Guerrieri Gonzaga, Monica Marabelli, Davide Bondavalli and the personnel of the Cogentech Cancer Genetic Test Laboratory. The MCCS was made possible by the contribution of many people, including the original investigators, the teams that recruited the participants and continue working on follow-up, and the many thousands of Melbourne residents who continue to participate in the study. We thank the coordinators, the research staff and especially the MMHS participants for their continued collaboration on research studies in breast cancer. The following are NBCS Collaborators: Anne-Lise Børresen-Dale (Prof. Em.), Kristine K. Sahlberg (PhD), Lars Ottestad (MD), Rolf Kåresen (Prof. Em.) Dr. Ellen Schlichting (MD), Marit Muri Holmen (MD), Toril Sauer (MD), Vilde Haakensen (MD), Olav Engebråten (MD), Bjørn Naume (MD), Alexander Fosså (MD), Cecile E. Kiserud (MD), Kristin V. Reinertsen (MD), Åslaug Helland (MD), Margit Riis (MD), Jürgen Geisler (MD), OSBREAC and Grethe I. Grenaker Alnæs (MSc). NBHS and SBCGS thank study participants and research staff for their contributions and commitment to the studies. For NHS and NHS2 the study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We would like to thank the participants and staff of the NHS and NHS2 for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. OBCS thanks Arja Jukkola-Vuorinen, Mervi Grip, Saila Kauppila, Meeri Otsukka, Leena Keskitalo and Kari Mononen for their contributions to this study. The OFBCR thanks Teresa Selander, Nayana Weerasooriya and Steve Gallinger. ORIGO thanks E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The LUMC survival data were retrieved from the Leiden hospital-based cancer registry system (ONCDOC) with the help of Dr. J. Molenaar. PBCS thanks Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The ethical approval for the POSH study is MREC /00/6/69, UKCRN ID: 1137. We thank staff in the Experimental Cancer Medicine Centre (ECMC) supported Faculty of Medicine Tissue Bank and the Faculty of Medicine DNA Banking resource. The RBCS thanks Jannet Blom, Saskia Pelders, Wendy J.C. Prager – van der Smissen, and the Erasmus MC Family Cancer Clinic. SBCS thanks Sue Higham, Helen Cramp, Dan Connley, Ian Brock, Sabapathy Balasubramanian and Malcolm W.R. Reed. We thank the SEARCH and EPIC teams. SZBCS thanks Ewa Putresza. UCIBCS thanks Irene Masunaka. UKBGS thanks Breast Cancer Now and the Institute of Cancer Research for support and funding of the Generations Study, and the study participants, study staff, and the doctors, nurses and other health care providers and health information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre.

BCAC is funded by the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and the PERSPECTIVE I&I project, funded by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie et de l'Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation. The EU Horizon 2020 Research and Innovation Programme funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. Additional funding for BCAC is provided via the Confluence project which is funded with intramural funds from the National Cancer Institute Intramural Research Program, National Institutes of Health.

Genotyping of the OncoArray was funded by the NIH Grant U19 CA148065, and Cancer UK Grant C1287/A16563 and the PERSPECTIVE project supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344) and, the Ministère de l'Économie, Science et Innovation du Québec through Genome Québec and the PSRSIIRI-701 grant, and the Quebec Breast Cancer Foundation. Funding for iCOGS came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, and Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The DRIVE Consortium was funded by U19 CA148065.

The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Senior Principal Research Fellow. M.C.S. is a NHMRC Senior Research Fellow. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]. The Australian Breast Cancer Tissue Bank (ABCTB) was supported by the National Health and Medical Research Council of Australia, The Cancer Institute NSW and the National Breast Cancer Foundation. The AHS study is supported by the intramural research program of the National Institutes of Health, the National Cancer Institute (grant number Z01-CP010119), and the National Institute of Environmental Health Sciences (grant number Z01-ES049030). The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breast Cancer Now and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). The BCEES was funded by the National Health and Medical Research Council, Australia and the Cancer Council Western Australia and acknowledges funding from the National Breast Cancer Foundation (JS). For the BCFR-NY this work was supported by grant UM1 CA164920 from the National Cancer Institute. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government or the BCFR. The BCINIS study is supported in part by the Breast Cancer Research Foundation (BCRF). For BIGGS, ES is supported by NIHR

Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, United Kingdom. IT is supported by the Oxford Biomedical Research Centre. The BREast Oncology GAlician Network (BREOGAN) is funded by Acción Estratégica de Salud del Instituto de Salud Carlos III (ISCIII) FIS PI12/02125/Cofinanciado FEDER, and ISCIII/PI17/00918/Cofinanciado FEDER; Acción Estratégica de Salud del Instituto de Salud Carlos III FIS Intrasalud (PI13/01136); Programa Grupos Emergentes, Cancer Genetics Unit, Instituto de Investigacion Biomedica Galicia Sur. Xerencia de Xestion Integrada de Vigo-SERGAS, Instituto de Salud Carlos III, Spain; Grant 10CSA012E, Consellería de Industria Programa Sectorial de Investigación Aplicada, PEME I + D e I + D Suma del Plan Gallego de Investigación, Desarrollo e Innovación Tecnológica de la Consellería de Industria de la Xunta de Galicia, Spain; Grant EC11-192. Fomento de la Investigación Clínica Independiente, Ministerio de Sanidad, Servicios Sociales e Igualdad, Spain; and Grant FEDER-Innterconecta. Ministerio de Economia y Competitividad, Xunta de Galicia, Spain. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). CBCS is funded by the Canadian Cancer Society (grant # 313404) and the Canadian Institutes of Health Research. CCGP is supported by funding from the University of Crete. The CECILE study was supported by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Agence Nationale de Sécurité Sanitaire, de l'Alimentation, de l'Environnement et du Travail (ANSES), Agence Nationale de la Recherche (ANR). The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council, and Herlev and Gentofte Hospital. The CNIO-BCS was supported by the Instituto de Salud Carlos III, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI12/00070). The American Cancer Society funds the creation, maintenance, and updating of the CPS-II cohort. The California Teachers Study and the research reported in this publication were supported by the National Cancer Institute of the National Institutes of Health under award number U01-CA199277; P30-CA033572; P30-CA023100; UM1-CA164917; and R01-CA077398. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health. The collection of cancer incidence data used in the California Teachers Study was supported by the California Department of Public Health pursuant to California Health and Safety Code Section 103885; Centers for Disease Control and Prevention’s National Program of Cancer Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute’s Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I awarded to the University of California, San Francisco, contract HHSN261201800015I awarded to the University of Southern California, and contract HHSN261201800009I awarded to the Public Health Institute. The opinions, findings, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the official views of the State of California, Department of Public Health, the National Cancer Institute, the National Institutes of Health, the Centers for Disease Control and Prevention or their Contractors and Subcontractors, or the Regents of the University of California, or any of its programs. The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by: Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l’Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF) (Germany); the Hellenic Health Foundation, the Stavros Niarchos Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS), PI13/00061 to Granada, PI13/01162 to EPIC-Murcia, Regional Governments of Andalucía, Asturias, Basque Country, Murcia and

Navarra, ISCIII RETIC (RD06/0020) (Spain); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (United Kingdom). The ESTHER study was supported by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). FHRISK is funded from NIHR grant PGfAR 0707-10031. The GC-HBOC (German Consortium of Hereditary Breast and Ovarian Cancer) is supported by the German Cancer Aid (grant no 110837, coordinator: Rita K. Schmutzler, Cologne). This work was also funded by the European Regional Development Fund and Free State of Saxony, Germany (LIFE - Leipzig Research Centre for Civilization Diseases, project numbers 713-241202, 713-241202, 14505/2470, 14575/2470). The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany. The GESBC was supported by the Deutsche Krebshilfe e. V. [70492] and the German Cancer Research Center (DKFZ). The HABCS study was supported by the Claudia von Schilling Foundation for Breast Cancer Research, by the Lower Saxonian Cancer Society, and by the Rudolf Bartling Foundation. The HEBCS was financially supported by the Helsinki University Hospital Research Fund, the Finnish Cancer Society, and the Sigrid Juselius Foundation. The HMBCS was supported by a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation. The HUBCS was supported by a grant from the German Federal Ministry of Research and Education (RUS08/017), B.M. was supported by grant 17-44-020498, 17-29-06014 of the Russian Foundation for Basic Research, D.P. was supported by grant 18-29-09129 of the Russian Foundation for Basic Research, E.K was supported by the program for support the bioresource collections №007-030164/2, and the study was performed as part of the assignment of the Ministry of Science and Higher Education of the Russian Federation (№AAAA-A16-116020350032-1). Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and Bert von Kantzows foundation. The KARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. The KBCP was financially supported by the special Government Funding (VTR) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, and by the strategic funding of the University of Eastern Finland. kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], Cancer Council Victoria, Queensland Cancer Fund, Cancer Council New South Wales, Cancer Council South Australia, The Cancer Foundation of Western Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow. LMBC is supported by the 'Stichting tegen Kanker'. DL is supported by the FWO. The MABCS study is funded by the Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", MASA. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419, 110826, 110828], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. The MASTOS study was supported by “Cyprus Research Promotion Foundation” grants 0104/13 and 0104/17, and the Cyprus Institute of Neurology and Genetics. MBCSG is supported by grants from the Italian Association for

Cancer Research (AIRC). The MCBCS was supported by the NIH grants R35CA253187, R01CA192393, R01CA116167, R01CA176785 a NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [P50CA116201], and the Breast Cancer Research Foundation. The Melbourne Collaborative Cohort Study (MCCS) cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database. The MEC was supported by NIH grants CA63464, CA54281, CA098758, CA132839 and CA164973. The MISS study is supported by funding from ERC-2011-294576 Advanced grant, Swedish Cancer Society, Swedish Research Council, Local hospital funds, Berta Kamprad Foundation, Gunnar Nilsson. The MMHS study was supported by NIH grants CA97396, CA128931, CA116201, CA140286 and CA177150. The work of MTLGEBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research; the Research Council of Norway grant 193387/V50 (to A-L Børresen-Dale and V.N. Kristensen) and grant 193387/H10 (to A-L Børresen-Dale and V.N. Kristensen), South Eastern Norway Health Authority (grant 39346 to A-L Børresen-Dale) and the Norwegian Cancer Society (to A-L Børresen-Dale and V.N. Kristensen). The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The Northern California Breast Cancer Family Registry (NC-BCFR) and Ontario Familial Breast Cancer Registry (OFBCR) were supported by grant U01CA164920 from the USA National Cancer Institute of the National Institutes of Health. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The Carolina Breast Cancer Study (NCBCS) was funded by Komen Foundation, the National Cancer Institute (P50 CA058223, U54 CA156733, U01 CA179715), and the North Carolina University Cancer Research Fund. The NHS was supported by NIH grants P01 CA87969, UM1 CA186107, and U19 CA148065. The NHS2 was supported by NIH grants UM1 CA176726 and U19 CA148065. The OBCS was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Center of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. Genotyping for PLCO was supported by the Intramural Research Program of the National Institutes of Health, NCI, Division of Cancer Epidemiology and Genetics. The PLCO is supported by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, National Institutes of Health. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. The SBCS was supported by Sheffield Experimental Cancer Medicine Centre and Breast Cancer Now Tissue Bank. SEARCH is funded by Cancer Research UK [C490/A10124, C490/A16561] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. The University of Cambridge has received salary support for

PDPP from the NHS in the East of England through the Clinical Academic Reserve. The Sister Study (SISTER) is supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005 and Z01-ES049033). The SMC is funded by the Swedish Cancer Foundation and the Swedish Research Council (VR 2017-00644) grant for the Swedish Infrastructure for Medical Population-based Life-course Environmental Research (SIMPLER). The SZBCS was supported by Grant PBZ_KBN_122/P05/2004 and the program of the Minister of Science and Higher Education under the name "Regional Initiative of Excellence" in 2019-2022 project number 002/RID/2018/19 amount of financing 12 000 000 PLN. The TNBCC was supported by: a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation. The UCIBCS component of this research was supported by the NIH [CA58860, CA92044] and the Lon V Smith Foundation [LVS39420]. The UKBGS is funded by Breast Cancer Now and the Institute of Cancer Research (ICR), London. ICR acknowledges NHS funding to the NIHR Biomedical Research Centre. The UKOPS study was funded by The Eve Appeal (The Oak Foundation) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The US3SS study was supported by Massachusetts (K.M.E., R01CA47305), Wisconsin (P.A.N., R01 CA47147) and New Hampshire (L.T.-E., R01CA69664) centers, and Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The USRT Study was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA.

NBCS Collaborators

Vessela N. Kristensen[115] , Kristine K. Sahlberg [116,117], Anne-Lise Børresen-Dale[75,118], Inger Torhild Gram[119], Olav Engebråten[75,117], Bjørn Naume[120], Jürgen Geisler[121], Grethe I. Grenaker Alnæs[115]

ABCTB Investigators

Christine Clarke[28], Jane Carpenter[122], Deborah Marsh[123], Rodney Scott[124], Robert Baxter[125], Desmond Yip [126, 127], Alison Davis[128, 129], Nirmala Pathmanathan[130, 131], Peter Simpson[132], Dinny Graham[122], Mythily Sachchithananthan[122]

kConFab/AOCS Investigators

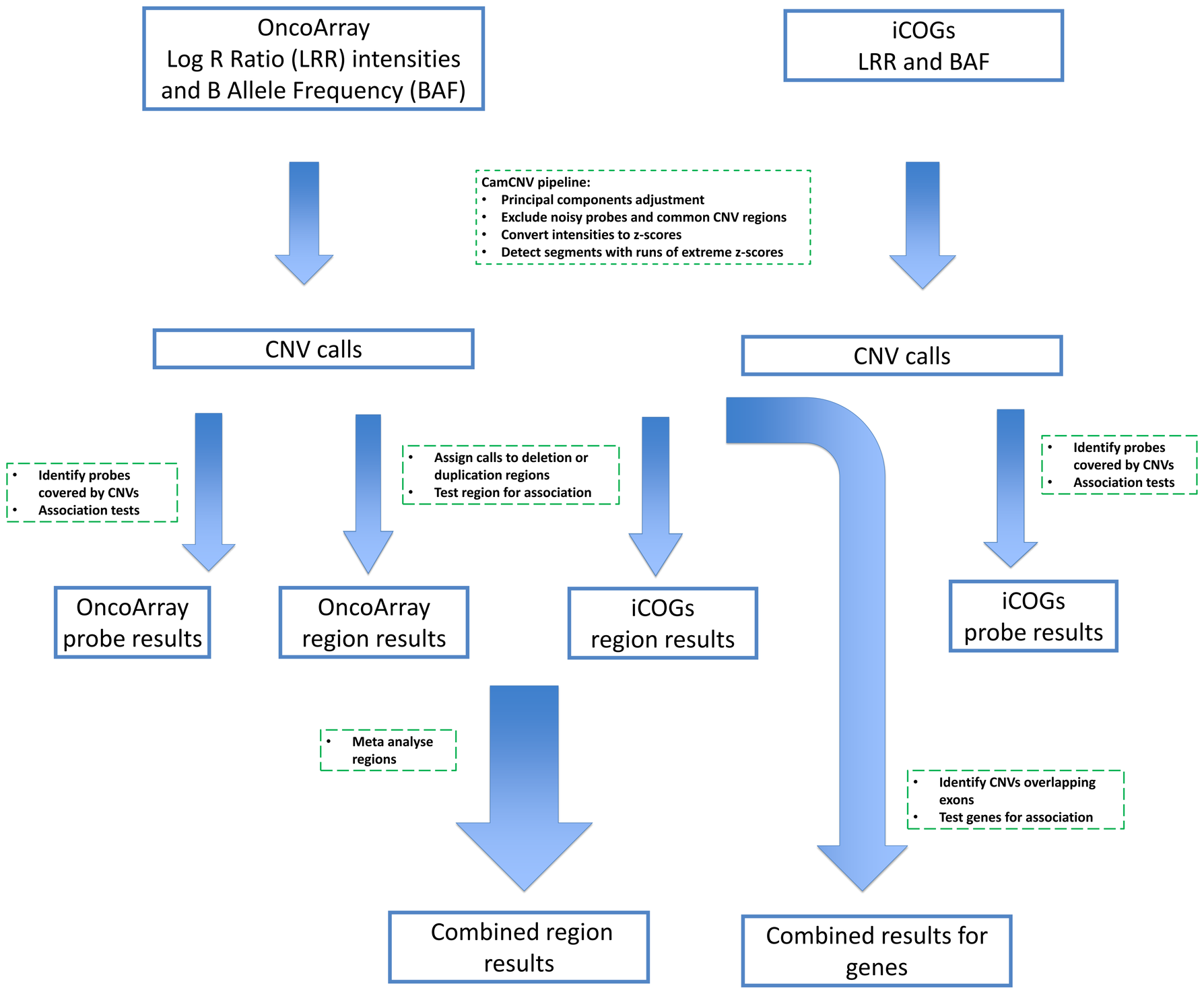
Ian Campbell[133], Georgia Chenevix-Trench[27], Anna de Fazio[134], Stephen Fox[133], Judy Kirk[135], Geoff Lindeman[136], Roger Milne[49-51], Melissa Southey[49,51,102], Amanda Spurdle[137], Heather Thorne[133]

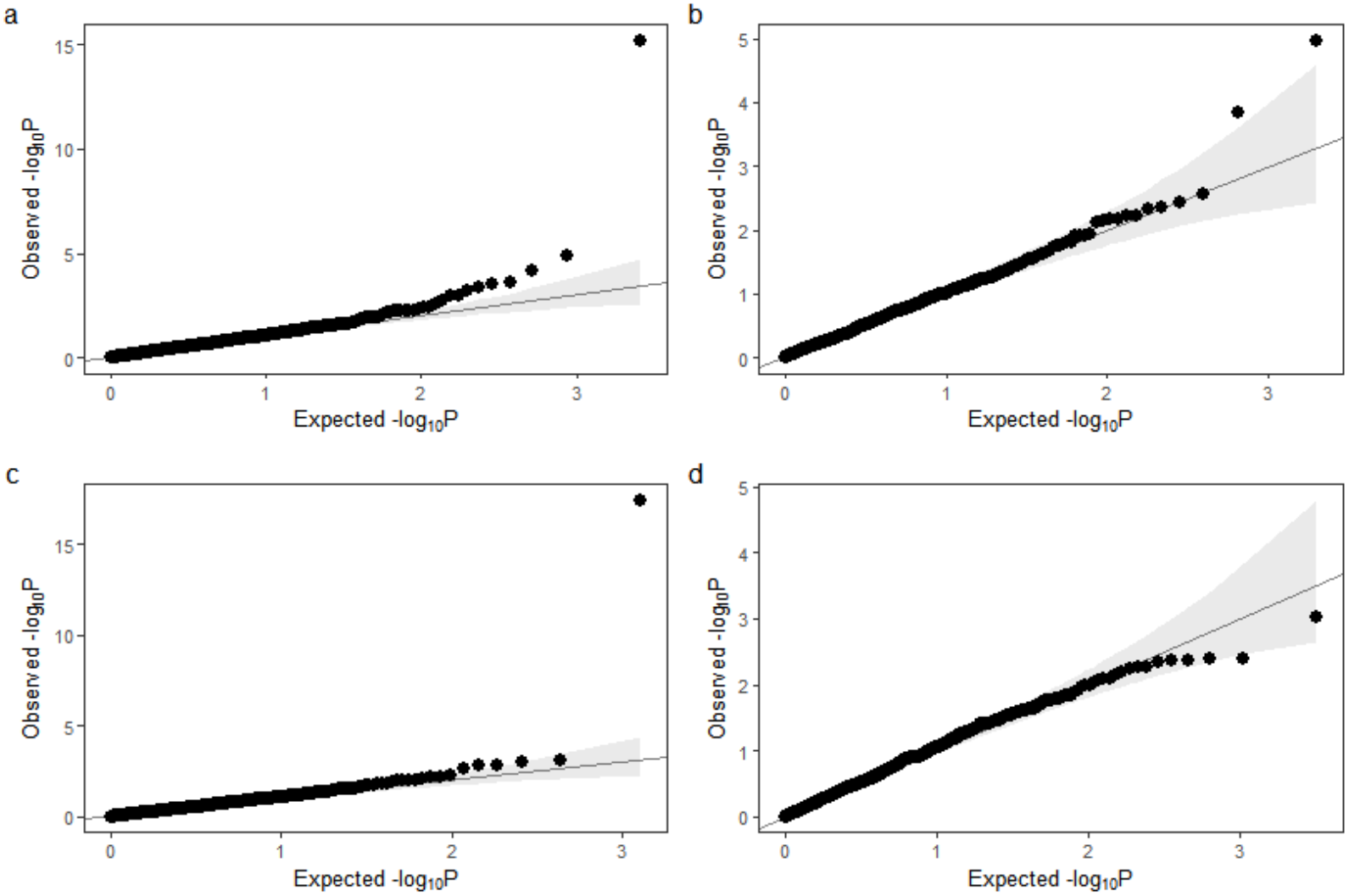
CTS Consortium

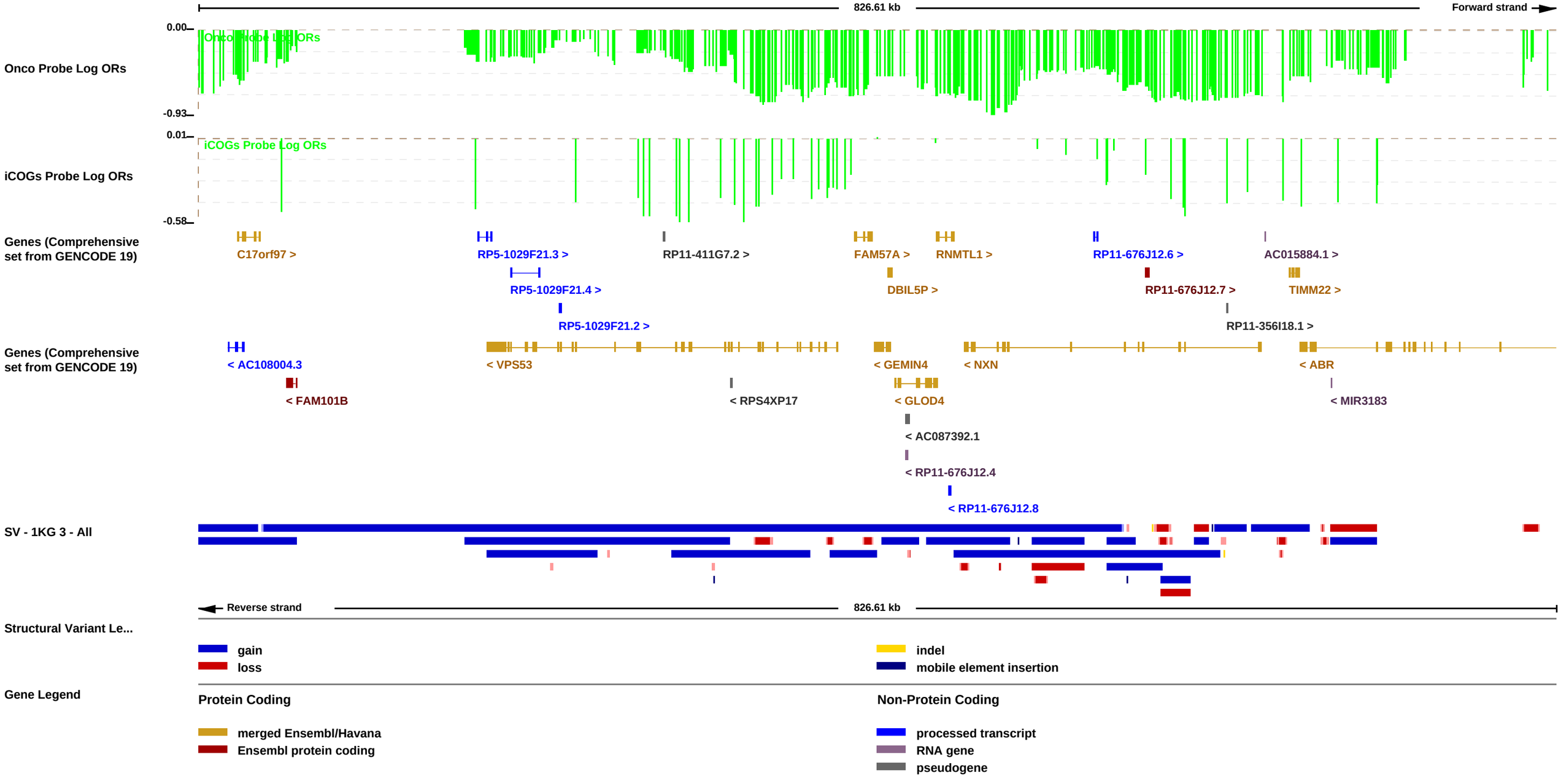
James Lacey[77], Elena Martinez[138]

[115] Department of Medical Genetics, Oslo University Hospital and University of Oslo, Norway
[116] Department of Research, Vestre Viken Hospital, Drammen, Norway

[117] Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital - Radiumhospitalet, Oslo, Norway
[118] Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway
[119] Department of Community Medicine, The Arctic University of Norway, Tromsø, Norway
[120] Department of Oncology, Division of Surgery and Cancer and Transplantation Medicine, Oslo University Hospital-Radiumhospitalet, Oslo, Norway
[121] Department of Oncology, Akershus University Hospital, Lørenskog, Norway
[122] The Westmead Institute for Medical Research, The University of Sydney, Sydney, NSW, Australia.
[123] University of Technology Sydney, Translational Oncology Group, School of Life Sciences, Faculty of Science, Ultimo, NSW, Australia
[124] School of Biomedical Sciences, University of Newcastle, Newcastle; Hunter Medical Research Institute and NSW Health Pathology North, Newcastle, Australia
[125] Kolling Institute of Medical Research, University of Sydney, St Leonards, NSW, Australia.
[126] Epigenetics & Transcription Laboratory Melanie Swan Memorial Translational Centre, Sci-Tech, University of Canberra, Canberra, Australia
[127] Department of Medical Oncology, The Canberra Hospital, Garran, ACT, Australia
[128] The Canberra Hospital, Garran, ACT, Australia
[129]The Australian National University, ACT, Australia
[130] Westmead Breast Cancer Institute, Western Sydney Local Health District, Westmead, New South Wales, Australia
[131] University of Sydney, Western Clinical School, Westmead, New South Wales, Australia
[132] UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Herston, QLD, Australia
[133] Peter MacCallum Cancer Centre, Melbourne, Australia
[134] Westmead Institute for Cancer Research, Sydney, Australia
[135] Dept. of Medicine, Westmead Hospital, Sydney, Australia
[136] Walter and Eliza Hall Institute, Melbourne, Australia
[137] Queensland Institute of Medical Research, Brisbane, Australia
[138] University of California, San Diego, CA, USA


OncoArray
Log R Ratio (LRR) intensities
and B Allele Frequency (BAF)
iCOGs
LRR and BAF
CamCNV pipeline:
• Principal components adjustment
• Exclude noisy probes and common CNV regions
• Convert intensities to z-scores
• Detect segments with runs of extreme z-scores
CNV calls
CNV calls
• Identify probes covered by CNVs
• Association tests
• Assign calls to deletion or duplication regions
• Test region for association
• Identify probes covered by CNVs
• Association tests
OncoArray probe results
OncoArray region results
iCOGs region results
iCOGs probe results
• Meta analyse regions
• Identify CNVs overlapping exons
• Test genes for association
Combined region results
Combined results for genes


a
b
c
d
Observed -log10P
Expected -log10P
0
1
2
3
5
10
15
4


826.61 kb
Forward strand
Onco Probe Log ORs
0.00
-0.93
iCOGs Probe Log ORs
0.01
-0.58
Genes (Comprehensive set from GENCODE 19)
C17orf97 >
RP5-1029F21.3 >
RP11-411G7.2 >
FAM57A >
RNMTL1 >
RP11-676J12.6 >
AC015884.1 >
RP5-1029F21.4 >
DBIL5P >
RP11-676J12.7 >
TIMM22 >
RP5-1029F21.2 >
RP11-356I18.1 >
Genes (Comprehensive set from GENCODE 19)
< AC108004.3
< VPS53
< GEMIN4
< NXN
< ABR
< FAM101B
< RPS4XP17
< GLOD4
< MIR3183
< AC087392.1
< RP11-676J12.4
< RP11-676J12.8
SV - 1KG 3 - All
Reverse strand
826.61 kb
Structural Variant Le...
gain
loss
indel
mobile element insertion
Gene Legend
Protein Coding
merged Ensembl/Havana
Ensembl protein coding
Non-Protein Coding
processed transcript
RNA gene
pseudogene